Analysis of ethical guidelines for AI systems

Guidelines with a dominant focus on general ethical principles

The European Group on Ethics in Science and New Technologies set out to create a common, internationally recognized ethical framework for the design, production, use and governance of AI. More than any other guideline, they urge us to be aware of autonomous systems, especially Lethal Autonomous Weapon Systems (LAWS), self-driving cars and autonomous software (including bots). The questions the group identified focus on the human control and responsibility over these systems, their explicability and if we can identify when we deal with autonomous systems.

Fairness, accountability and transparency in Machine Learning (FAT/ML) focuses on "algorithmic systems", and it aims to help developers and product managers design and implement systems that are publicly accountable. Along with the ethical principles, they devised a social impact assessment, consisting of guiding questions and initial steps to take. These include responsibility, explainability, accuracy, auditability and fairness. These are more concrete considerations that can create ground for more specific plans, but they are far from complete.

AI4People's ethical framework identified a set of ethical principles taken from bioethics. They adapted a fifth principle, particularly tailored to address artificial intelligence: explicability. According to their principles, we need to strive for transparency firstly by appointing responsible actors and also by explaining technical details and decisions that were made. This is followed by a set of concrete recommendations that follow these ethical principles. The purpose of them is to support responsible governance and an enabling environment for ethically aligned AI.

Interactive guidelines

Algo.rules is a set of criteria that readers can click through, developed by Bertelsmann Stiftung & i.RightsLab. It was written for every stakeholder that is involved during the AI system lifecycle. Although the explanations are short, they provide concrete recommendations in line with the identified principles. Among other things, they ensure that responsibilities are defined, transparency is achieved by documentation, that the algorithmic system is robust, that potential harm can be mitigated and that external auditing is possible.

The Center for Democracy and Technology (CDT) considered similar ethical principles, but instead of listing them, they created an interactive tool encompassing the whole AI system lifecycle. The questions are guided by ethical considerations. As an example, they ask "have the tools you are using been associated with biased products?" or "can you determine metrics that demonstrate the reliability of your model?" The questions can guide creators of the algorithms, but also those in the public sector, to ask relevant questions.

Guidelines that build on existing legal frameworks

OECD issued a set of recommendations for all stakeholders that are involved in the AI system lifecycle, and also for governments to foster an environment for responsible AI development. They adhere to certain ethical principles such as sustainable development, human-centered values, fairness, transparency, robustness and accountability. These are short recommendations to be adapted by governments, but they do not offer any concrete tools to put these into action.

The Alan Turing Institute's guide was written in partnership with Government Digital Service (GDS) and Office for Artificial Intelligence (OAI) of the UK government. It targets workers in the public sector, to help governments innovate with data-intensive technologies. The guideline evolves from more abstract and ethical principles (SUM values) that serve as guiding principles, to the more concrete actionable principles (FAST principles). The end goal is to implement the ethical and actionable principles through a process based government (PBG) framework. This includes assigning responsibilities within the team, ensuring transparency and assessing the impact more than once during the AI system lifecycle . The guide advises to use checklists, in order to secure a responsible delivery through human centered protocols and practices.

The European Union's High Level Expert Group on Artificial Intelligence's guide was created to strengthen ethical and legal principles according to European values. It builds on EU law, and departs from more abstract ethical principles to more actionable ones. The guide contains 7 requirements that are needed to achieve these ethical principles. For example, regarding accountability, this guide identifies auditability (openly available information about business models and intellectual property related to the AI system), minimisation, and reporting of negative impacts (use of impact assessments). The guide also includes an impact assessment which is a collaboration of different stakeholders, including a set of relevant questions that civil servants and other involved parties can utilize. The creators of this guide invite stakeholders to pilot the assessment list and provide feedback on its implementability and completeness.

IEEE pays special attention to driving ethically aligned design. It offers guidance for standards, legislation, regulation, guided by ethical principles. In the form of a lengthy text, they build on more abstract ethical principles such as well-being and awareness of misuse, in order to derive more concrete recommendations from them.

Finally, the Canadian government issued a directive on automated decision making, which demonstrates its efforts to use algorithmic systems implemented in administrative decisions in more accountable and transparent ways. The government issues a responsible person who ensures that prior to the production of any autonomous decision making system, they carry out an algorithmic impact assessment. The impact assessment asks particular questions that they need to answer and it assigns the particular project to a certain level. These levels indicate a ladder of social impact, pointing to varying levels of follow-up actions.

Author: Petra Biró